Final Report

Gender Bias in Natural Language Processing: BioCorpus-5, A Preliminary Multilingual Gender-Balanced Corpus of In-domain Wikipedia Biographies

A Degree Thesis
Submitted to the Faculty of the
Escola Tècnica d'Enginyeria de Telecomunicació de Barcelona
Universitat Politècnica de Catalunya
by

Kim Jung, Jae Hyouk

Advisor:

Ruiz Costa-jussà, Marta

Barcelona, January 2019

Abstract

In natural language processing and the blind application of machine learning reflect social biases and stereotypes in training data. In this project, we develop a corpus for future analysis applications of this bias. The corpus uses the data extracted by a tool called Wiki-Tailor which helps to obtain multilingual biographies from Wikipedia.

The extracted multilingual corpus of biographies based on actors, linguists and physicists is analyzed, and it is balanced in gender in five languages: Spanish, Catalan, French, English and German. For that purpose, it is necessary to create a semi-automatic software divided into two parts.

On the one hand, a manual alignment of the text of each biography is carried out in order to obtain five text files for each author where the information of each line is parallel for each language. To do this, information is extracted in the five different languages that coincide in content and are aligned and/or translated to achieve the correct format to do a future revision to ensure that the result is as natural as possible. On the other hand, each file is formatted in each language parallelized in xml. The xml data enters each author's information (identifier, language, genre, etc.) and is presented in a single text file to make the system simpler and more useful to process.

Finally, statistics are obtained from the corpus created so it can be used in future automatic natural language processing or machine learning applications which require multilingual parallel corpus either at the level of sentence or document.

Resum

En el processament del llenguatge natural (NLP), els sistemes neurals de traducció automàtica i l'aplicació de l'aprenentatge automàtic reflecteixen bias i estereotips socials a l'entrenament de dades. En aquest projecte es crea un corpus amb futures aplicacions d'anàlisi d'aquest bias a partir de les dades extretes d'una eina anomenada Wiki-Tailor, que ajuda a obtenir biografies multilingües de Wikipedia.

Aquest corpus de biografies multilingües extretes centrades en actors, físics i lingüistes és analitzat i balancejat en cinc idiomes diferents: castellà, català, francès, anglès i alemany. Per la seva realització, és necessària la creació d'un software semiautomàtic dividit en dos parts.

En primer lloc, es realitza un alineament manual del text de cada biografia per obtenir com a resultat cinc arxius de text per a cada autor, on la informació de cada línia és paral·lela per a cada idioma. Per fer això s'extreuen les informacions en els cinc idiomes diferents que coincideixen en contingut i s'alinien i/o tradueixen per tal de tenir el format correcte per posteriorment revisar que el resultat sigui el més natural possible. En segon lloc, s'utilitzen les dades en xml per marcar la informació paral·lela de cada autor (identificador, idioma, gènere, etc.) i es presenten en un fitxer de text únic perquè el sistema sigui més senzill i útil de processar.

Finalment s'obtenen estadístiques del corpus creat per poder ser utilitzat en futures aplicacions de processament automàtic del llenguatge natural o d'aprenentatge automàtic que requereixin corpus paral·lel multilingüe, sigui a nivell d'oració o de document.

Resumen

En el procesado del lenguaje natural (NLP), los sistemas neurales de traducción automática y la aplicación ciega del aprendizaje automático reflejan bias en los datos de entrenamiento. En este proyecto se crea un corpus con futuras aplicaciones de análisis de este bias a partir de los datos extraídos por una herramienta llamada Wiki-Tailor, que ayuda a obtener biografías multilingües de Wikipedia.

Este corpus de biografías multilingües extraídas centrada en actores, físicos y lingüistas es analizado y balanceado en cinco idiomas: castellano, catalán, francés, inglés y alemán. Para ello, es necesaria la creación de un software semiautomático dividido en dos partes.

En primer lugar, se realiza una alineación manual del texto de cada biografía para obtener como resultado cinco archivos de texto para cada autor donde la información de cada línea es paralela para cada idioma. Para ello, se extraen las informaciones en los cinco idiomas diferentes que coinciden en contenido y se alinean y/o se traducen para conseguir el formato correcto para posteriormente revisar que el resultado sea lo más natural posible. En segundo lugar, se da formato a cada archivo en cada idioma paralelizado en xml. Los datos xml entran la información de cada autor (identificador, idioma, género, etc.) y se presentan en un archivo de texto único para que el sistema sea más sencillo y útil de procesar.

Finalmente se obtienen estadísticas del corpus creado para que pueda ser utilizado en futuras aplicaciones de procesamiento automático del lenguaje natural o de aprendizaje automático que requieran corpus paralelo multilingüe, ya sea a nivel de oración o de documento.

Acknowledgements

First, I want to thank my tutor Marta Ruiz Costa-Jussà and Cristina España-Bonet for letting me be part of this project and motivate me to continue investigating about the interesting world of Artificial Intelligence. Also, I want to express my gratitude to Cristina Abad Moya for the immense patience she had with me, listening and supporting me being at my side whenever I needed it.

On the other hand, I want to acknowledge my supervisor Hae Joon Jung's advice and help during my stay in Incheon National University helping me with all the problems I had.

And finally, I want to thank my parents for all the dedication and effort they have had with me to bring me here and for being by my side supporting me in difficult times.

Revision history and approval record

Revision	**Date**	**Purpose**
0	08/01/2019	Document creation
1	24/01/2019	Document revision

DOCUMENT DISTRIBUTION LIST

Name	**e-mail**
Jae Hyouk Kim Jung	alexkimbcn@gmail.com
Marta Ruiz Costa-jussà	marta.ruiz@upc.edu

Written by:		Reviewed and approved by:	
Date	08/01/2019	Date	24/01/2019
Name	Jae Hyouk Kim Jung	Name	Marta Ruiz Costa-jussà
Position	Project Author	Position	Project Supervisor

Table of contents

1 Introduction

Natural language processing is a branch of machine learning that has grown in importance in recent years. This type of machine learning application provides state-of-the-art models for tasks such as language modeling or machine translation. This project focuses on the bias created in this kind of application, machine translation. This is because the data in these models are trained by human language texts, so a natural question is whether they exhibit biases based on gender or other characteristics, and, if so, how should this bias be mitigated.

To apply these tasks, the creation of a corpus is required. In this project focused on the mitigation of gender bias, it has been decided to create a corpus with biographies balanced in gender in at least five languages. The question is to be able to compare biographies of different masculine and feminine gender authors in order to study this bias that is created by decanting one gender or another. For this, the help of a project that extracts the biographies from Wikipedia in different languages is used and with it it will be possible to create a semiautomatic corpus. To be able to optimize the creation of this corpus requires a software that automates the formatting of all files that make up the corpus of all authors in all languages, so the software is programmed to facilitate the task of work for this large amount of data, which is specified later.

1.1 Statement of purpose

The main objective of the project is to create, from a multilingual database extracted from Wikipedia, a xml balanced corpus data in five different languages for future automatic learning bias analysis applications.

The files extracted from Wikipedia do not share the same information between different languages of the same author, which complicates the parallelization of the texts. For this reason, first, a general analysis of the texts in the five different languages is carried out, extracting the sentences that they share among them (most of the times the first phrases of the author's birth description). In order to ensure that the corpus is not scarce of information, a reference language is used, in this case French, to translate the information in the other languages using an automatic translator to structure the body of each author and ensure that the information remains parallel.

Because of that, a post-processing of the corpus is required to revise the translator languages so that the result is as natural as possible.

Having a large extracted database that complicates working manually, it is chosen to create a software that maps each text file in each author's language so that it can be converted and

xml data can be extracted with the information from each parallel file.

Finally, final statistics are extracted from the corpus created with future proposals for automatic learning applications where this corpus can be used.

1.2 Requirements and specifications

To obtain Wikipedia's database focused on female and male physicists, linguists and actors, WikiTailor tool is required. Once the multilingual database has been extracted, a software using Visual Basic for Applications is created to obtain the corpus in xml.

This language is chosen for the simple reason of the need to have the information of each sentence of each parallel text file in the five languages. For this reason, to be able to work orderly and without character problems in any of the five languages, the Unicode UTF-8 coding is used in Excel, and therefore its application in Visual Basic to apply the software in each text file.

All the software is created by Microsoft Visual Basic for Applications 7.1 Version 1087 using a Microsoft Surface with Intel® processor Core™ i7 7660U CPU @ 2.50GHz, 64-bit operating system and 8.00 GB of RAM.

1.3 Methods and procedures

The main idea of the project is proposed by Dr. Marta Ruiz Costa-Jussà together with Cristina España-Bonet, a researcher on Natural Language Processing at University of Saarland. This thesis uses as starting point the automatically multilingual in-domain corpora from Wikipedia from actors, physicists and linguists extracted with the improved and extended tool, Wiki-Tailor.

The main core of the thesis is to parallelise at the level of sentence this corpus focusing on obtaining a gender-balanced resource. Once the corpus is created, a software is used to obtain statistics of the final balanced corpus and use the results for inspirations related to different experiments such as classifications of sentences by genre, being able to extract conclusions comparing the different languages.

1.4 Work Plan

This project was structured with the following work packages and Gantt diagram.

1.4.1 Work Packages

- WP 1: Project propose and work plan
- WP 2: BBDD analysis (4 languages)
- WP 3: Information research
- WP 4: Corpus creation
- WP 5: Project development and upgrade
- WP 6: Results and conclusions
- WP 7: Final presentation

1.4.2 Gantt Diagram

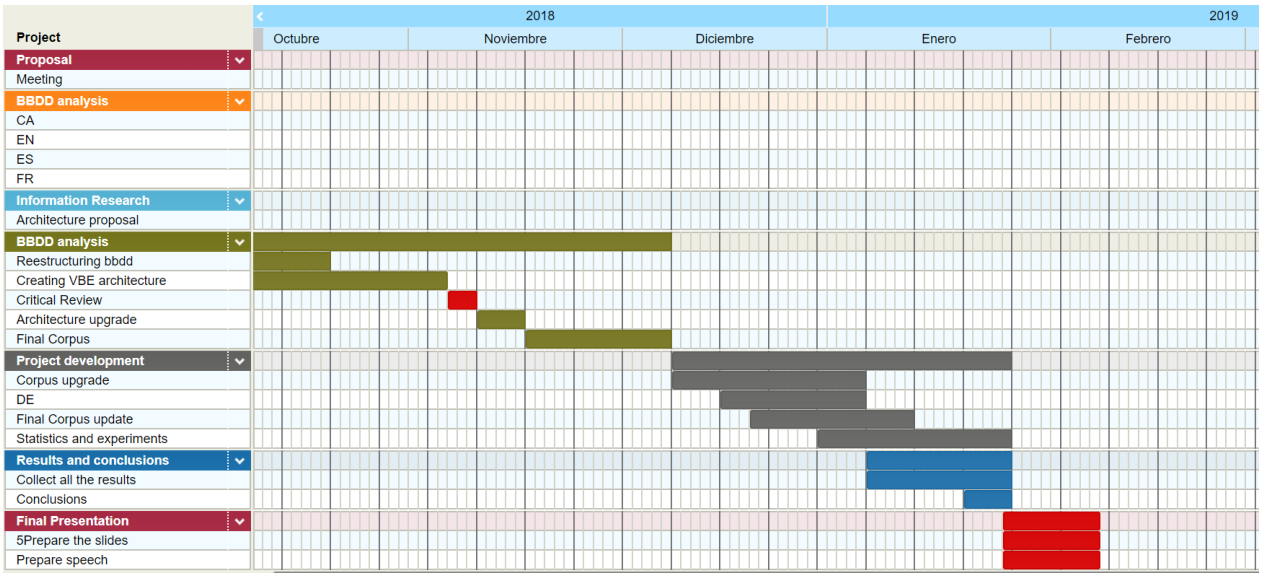

Figure 1.1: Gantt Diagram.

2 Background

This chapter explains the theoretical background that holds this project. First, this section defines the area which studies and creates corpus to automatic learning applications, Natural Language Processing. Then, this section provides a global vision of the context which the corpus created can work and goes deeper in the explanation of the specific software architectures chosen.

2.1 Machine Learning

Machine Learning in the context of text analytics is a set of statistical techniques for identifying parts of speech, entities, and other aspects of text. All the learning algorithms require a learning phase at which, an objective function is defined as a metric to optimize in order to get a reference of how well our model fits to the problem.

Then, the algorithm iterates through the training set looking for the optimization of the metric. It is important to have three disjoint sets of samples in machine learning algorithms: training, validation and test set. The training set is used as examples for the objective function optimization. A validation set is required when it is necessary to compute the optimal parameters of an algorithm. Finally, the test set is used to test how well the algorithm has learned and generalized the problem.

The techniques can be expressed as a model that is then applied to other text, also known as supervised machine learning. It also could be a set of algorithms that work across large sets of data to extract meaning, which is known as unsupervised machine learning. Difference lies in, if during the training process, training samples are labeled with information of the class they belong or conversely there is no additional information and is the system who must determine which class they belong to.

These algorithms reduce the human intervention at the time of defining rules or patterns to the systems, letting to them to extract that information. Their true potential has been possible thanks to recent computing capacity improvements and the availability of big data bases.

2.2 Natural Language Processing (NLP)

Natural Language Processing (NLP) is the part of Artificial Intelligence (AI) that deals with how a computer can be programmed to understand, process, and generate a language like a human.

NLP works through automatic learning (ML). Automated learning systems store words and the ways in which they are combined just like any other form of data. Phrases, sentences and sometimes complete books are processed according to grammatical rules, people's real-life linguistic habits, or both. The computer then uses this information to find patterns and extrapolate what comes next.

Although machines are not yet capable of equating a human in language recognition, this does not mean that they are not useful for various applications involving the ability to process and, in some cases, understand language. Some of these applications are already common, such as spam detection filters or spell checkers, but also more complex applications that have not yet been fully resolved, such as digital assistants, automatic translation systems or automated answers to questions.

The main feature of natural language that makes it difficult to automate its processing is ambiguity, something that the human brain usually manages to deal with. When interpreting a sentence, the human being evokes a whole series of personal and contextual experiences that impregnate it with meaning, which is very complex to model programmatically.

Another important challenge facing language processing is the wide variety of languages with different grammatical rules and regional variations for the same language. This causes a lack of generality in the proposed solutions and therefore many times the algorithms must be adapted specifically for each language or work only for some of them.

The first algorithms used to process the language were based on rules, but given the complexity of the language, the set of rules needed to model it tends to grow disproportionately, in addition to the fact that such rules are usually manually coded and require special cases for each possible interpretation of words.

Subsequently, the strategy that has been used to solve the problem in a pragmatic way is the use of probabilistic language models, which use large quantities of texts called corpus, which are processed to serve as examples of input to automatic learning algorithms. The idea is that these models somehow capture the frequency, order and ideally the semantics of words. In some algorithms and in some other applications it is common to use the word embeddings technique, in which words are represented as vectors of real numbers on which it is possible to carry out operations with some surprising results.

Normally these vectors are obtained by training a neural network that takes as input a large corpus of text and produces a vectoral representation for each word that captures the co-occurrence with the other words present in the corpus.

2.3 Neural Machine Translation (NMT)

Machine translation (MT) is automated translation. It is the process by which computer software is used to translate a text from one natural language (such as English) to another (such as Spanish).

When processing any translation, human or automatic, the meaning of the text in the original language must be fully restored in the target language. Although it may seem simple, it is much more complex. Translation is not a mere substitution of one word for another. A translator must interpret and analyze all the elements of the text and know how one word influences another. This requires extensive knowledge of grammar, syntax (sentence structure), semantics (meanings), etc., of the source and target languages, as well as familiarity with each specific region.

Both human and machine translation have their own challenges. For example, two individual translators cannot produce identical translations of the same text in the same language pair, and several rounds of proofreading may be required to achieve customer satisfaction. But the biggest challenge lies in how to produce quality translations that can be published using machine translation.

2.4 Visual Basic for Applications

Microsoft VBA (Visual Basic for Applications) is the Microsoft Visual Basic macro language used to program Windows applications and included in various Microsoft applications. VBA allows users and programmers to extend the program functionality of the Microsoft Office suite.

Its main utility is to automate everyday tasks, as well as create applications and database services for the desktop. It allows access to the functionalities of an event-oriented language with access to the Windows API.

Analyzing all the capacities that provides Microsoft Office and its programmer Visual Basic for Applications, it is chosen because it provides a more organized way to work with the data and modify massively large amounts of files.

3 Methodology

As previously started, the project has been divided in three parts along the semester, being the first part the analysis of the provided database from Wiki-Tailor, the second one the preparation of all the text files with the information in all the languages parallel and then the last part the application of the software to all the files to get the final corpus. In the following sections it will be explained how the extracted data-base was analysed and prepared and how the software was prepared to obtain the corpus semi-automatically.

All the code of the software has been written in VBA and Unicode UTF 8 codification has been used to work with all the files in the different languages to avoid strange character problems.

3.1 Wiki-Tailor

Wiki-Tailor[10] is a tool for extracting in-domain corpora from Wikipedia. It helped to extract the data from Wikipedia obtaining multilingual in-domain corpora of actors, physicists and linguists from Wikipedia. A domain is defined as an existing category in Wikipedia and the articles belonging to that domain are extracted even if they are not tagged as such.

Why Wikipedia? Wikipedia is probably the biggest crowd-sourced information platform with a built-in review process and as many languages as its users want it to be. It comes with a consistently maintained categorisation, so the categories plus text itself are classes in natural language processing (NLP). This last feature is the one that has made chosen Wikipedia as a source of information to extract the database that will be used to create the corpus of this project. It is a relatively reliable source, in different languages in which it is possible to extract biographies of many authors to perform gender studies using Wiki-Tailor, that performs the cleaning and sentence splitting of the input text, extracting only the title and the main body of the article. However, Wikipedia articles in different languages are not translations one of the other, but in most occasions the first and second paragraphs of an article are similar among languages. It is going to start by considering these cases for manually creating a sentence-aligned corpus. Cultural differences might widen the differences in the texts, so the initial set only includes European languages.

As stated above, the corpus will be used in machine translation evaluation so that at least 2,000 parallel sentences in the five languages will be extracted. The corpus will have a document structure where each document will be tagged with the ID of the original Wikipedia article, the language, the domain, and the gender of the person is referring to, because this structure allows to split the corpus in different subsets and evaluate domain-specific

translations.

3.2 Corpus creation

A linguistic corpus is a broad and structured set of real-life examples of language use. These examples can be texts (the most common), or oral samples (generally transcribed). In our case, a linguistic corpus, is a relatively large set of texts, created independently of their possible forms or uses. In other words, in terms of its structure, variety and complexity, a corpus must reflect a language as accurately as possible; in terms of its use, make sure that its representation is real.

In linguistics and NLP, corpus refers to a collection of texts. Such collections may be formed of a single language of texts or can span multiple languages. Corpora are generally solely used for statistical linguistic analysis and hypothesis testing.

The idea is that they represent language in the best possible way so that NLP models can learn the patterns necessary to understand the language. One of the most popular corpora that is usually used to train models is the one that includes the information extracted from Wikipedia. Wikipedia is a rich source of well-organized textual data and is a nice place to build a corpus because is conveniently available online and it is also a vast collection of knowledge.

As stated above, after the extraction of this corpus, there are three files (one per domain) with the name of the persons and ID of the file with the biography in 6 languages. The name of the file specifies the order of the languages in the columns and then, there is a folder per language with the text itself. It will be useful to be able to map all the files using the software that will automate the creation of the corpus.

The problem with the creation of this corpus is the fact that biographies of the same author in different languages do not share the same information. This causes that the most complicated task of the corpus creation is the parallelization of the information. Seeing the difficulty of analyzing the five files of each author to choose and modify so that the information is aligned, a mechanism is chosen that will require a post-processing of the corpus outside this project. The first step is filtering as much of the information shared in the five languages, which are usually the author's introductory phrases, and then, choosing a reference language, try to create the corpus structure by automatically translating into the remaining languages. Evidently, this will require a thorough post-processing review to modify and correct language errors that cause unnatural language.

This has been done by seeing the large amount of data and information and the difficult

analysis of five different languages without dominating them all perfectly. It has been seen that by performing this mechanics, the work of parallelization of information is more optimal in time, although later more time has to be devoted to post-processing, the corpus structure will already be implemented.

3.2.1 Representativeness and balance

A corpus could be conceived simply as a group of texts. However, a group of texts acquires corpus status when it is systematically compiled under certain parameters, which implies a fundamental methodological aspect in corpus studies. A corpus has been defined as a set or body of authentic texts in electronic format that are systematically selected and grouped under certain criteria for linguistic analysis. Three fundamental characteristics emerge from this definition. First, a corpus is structured and planned, as opposed to a set of texts without a particular order or logic. Second, a corpus is a set of authentic texts created with a real communicative purpose. Third, a corpus is a sample designed to represent a whole, such as a particular dialect. As can be observed in these three characteristics, the systematicity that characterizes the structure of a corpus allows conclusions to be reached and generalizations to be made about linguistic patterns based on empirical evidence of the language as it is used in real situations.

The systematic preparation of a corpus and the precision of its internal structure are very important aspects that determine the quality of research results. For this reason, the structure of the corpus must be designed by clearly defining the variety of language, genres and registers to be studied, and determining the procedures to be used for the selection of texts. Once the genres and records are defined, it is important to determine the internal structure of a corpus based on certain criteria so that the results are reliable and generalizable.

Two concepts, representativeness and balance, are fundamental for the construction of a corpus. As mentioned above, a corpus tries to be representative of a whole, therefore it must have similar characteristics to the language it represents. Representativeness is defined as the degree to which a sample includes the totality of variables of a population. On the other hand, a corpus is considered balanced when it is structured in different sections that contain equal proportions of certain variables, such as number of words. For this reason, there is currently little variety of balanced corpus, since the representativeness and balance of a corpus are not always objectives that can be fully defined and achieved. In this project, the corpus is balanced according to gender and according to whether the author is a linguist, physicist or actor, so that in the end a balanced corpus will be obtained in order to be able to carry out studies with the same proportion of both female and male authors.

3.3 VBA semi-automatic software

After the extraction of the database of actors, physicists and linguists in six different languages of both genders, it is necessary to find a way as automated as possible in order to obtain the corpus dedicating the least possible time and with the maximum accuracy as far as information in different languages.

The database consists of numerous text files with random identifiers and a text file with the id-author-language relationship to each of the corresponding text files. After analyzing the information provided and the order to follow, it is decided to work in Microsoft Excel platform and its programmer Visual Basic for Applications to be able to work and modify large amounts of files. (Figure 3.1).

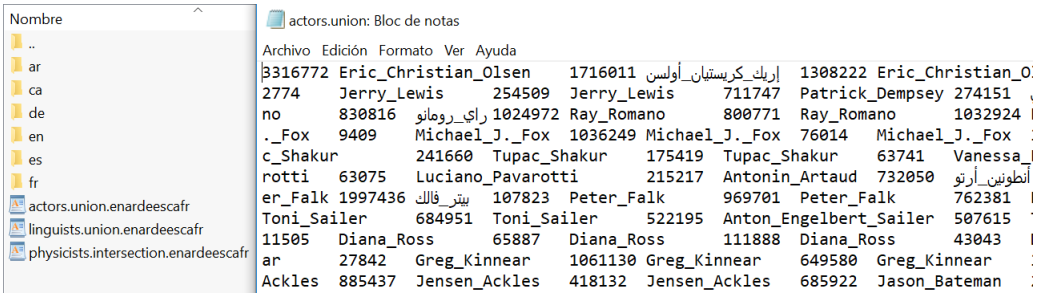

Figure 3.1: Extracted files from Wikipedia.

The main task at the time of creating the corpus is to have to parallelize all the text files of each language to obtain the same information of each author in each language. The main problems that are appreciated in the first instance are the following:

1. The files in each language of each author do not share any relationship. Only a provided text file helps to know which ID of which file belongs to the same author in each language.
2. The information of the same author in different languages is different, i.e. an author does not share the same information in any language.
3. Arabic language is discarded due to its difficulty of comprehension and the probability of increasing the translation error.

3.3.1 Renaming the archives

To solve the first problem of difficulty of relating the large amount of data by ID, first the correct encoding of opening files has been chosen so that they do not contain broken characters. Having the ID-author relationship in Excel provided by the Wiki-Tailor tool, it

is decided to modify all the files so that they appear by 'name'+'language' so that, when sorted alphabetically in the folder, they appear grouped and thus be able to open them more easily to modify them. In this way we save the task of searching for ID in the folder one by one and open them to modify the information. To massively modify all the files, a software is created in VBA that performs this work.(Figure 3.2).

6454	2322395.fr.txt	Brian_P._Schmidt_FR.txt
6455	476967.fr.txt	George_Stoney_FR.txt
6456	3022841.fr.txt	William_Rankine_FR.txt
6457	1308222.de.txt	Eric_Christian_Olsen_DE.txt
6458	206377.de.txt	Jonathan_Rhys_Meyers_DE.txt
6459	4626.de.txt	Schauspieler_DE.txt

Figure 3.2: Original names and modified names.

The software scrolls through the entire folder of files mapping by name of the first column and modifying the name that corresponds to the second column.

3.3.2 Parallelization of texts

It is needed to reformat each text file such that, for each author's language, it contains the same information on each line. To do this, taking advantage of the modification made before, the files are compared five by five (to open an author in five different languages) adding and/or removing information to be parallel in all languages.

To do this, the five files are opened in different languages in order to extract the information that matches them. In order to create a corpus with enough information, simple sentences are chosen to translate them by hand or using an automatic translator, to obtain parallel information. The reason for translating in this way is to express enough information in the five different languages to be able to create the structure of the corpus. For this purpose, the file in French has been chosen and sentences have been chosen to be translated by means of an automatic translator in the different languages. After that a post-process of the corpus of Spanish, Catalan, German and English is required to confirm the naturalness of the language and to correct translation inconsistencies.

Once it is assured that the information is parallel, the type of file is chosen that describes the data with greater precision in a delimited way that the characters (points or tabulations) separate the fields. In this way it is possible to have each author's file with its segments separated in Excel rows, so applying the following software to obtain xml data will be easier.

3.3.3 Software application

Once having all the files of each author with the parallel information, the next step is to apply the software. This software has the following features that help to obtain the final results, the following format of xml features:

```
1 <doc docid="Kathy_Bates_DE" wpid="171283" language="DE" topic="actors" genre="female">
2 <title> Kathy Bates </title>
3 <seg id="1">Kathy Bates</seg>
  <seg id="2">Kathy Bates ist eine Theaterschauspielerin, ihre erste Leidenschaft.</seg>
  <seg id="3">Sie trat 1971 zum ersten Mal im Kino in Taking Off von Miloš Forman auf.</seg>
4 </doc>
```

Figure 3.3: xml data.

- The directory where the files to be modified are located is indicated. The software scrolls through the entire folder and makes the modifications to each file.
- The 'docid' of point number 1, simply takes the name of the file, which distinguishes it with the 'wpid', which is the real identifier of the article in Wikipedia, so if someone wants to access automatically just use this id and not the title with all the encoding problems that involve.
- The 'language' category of point number 1, takes advantage of the renaming made in the first part of the software and as all files are formed by 'NAME'+'LANGUAGE (ES, CA, DE, FR, EN), and simply takes the last two values of the string of the name and prints the corresponding language.
- The 'topic' is automatically printed by relating the software with the three files provided by Wiki-Tailor that separates each author according to physical, linguistic or actor, so it simply maps the real wpid of each file and prints the type of author that is according to which file has been mapped.
- The 'genre' category is the part that has had to be done manually because there was no file that referenced this feature, so it has been separated manually by folders, on one hand males and females on the other, and the software has been executed separately changing the genre manually.
- The 'segment number' has been added by simply concatenating each line according to the ID of the segment and finally the end of the document is added when it finds the last line of the file.
- The 'html entities' have been modified to their standard coding that the parsers later recognize, according to the following table. In the software, each character has simply

been replaced by its entity name.

Result	Description	Entity Name	Entity Number
<	less than	&lt;	&#60;
>	greater than	&gt;	&#62;
&	ampersand	&amp;	&#38;
"	double quotation mark	&quot;	&#34;
'	single quotation mark (apostrophe)	&apos;	&#39;

Figure 3.4: Standard coding of html entities.

- Finally, the directory where want to extract all these files that have undergone these modifications is indicated recovering the real wpid of each one.

3.4 Balanced corpus

Once all the files with the software are extracted after being analyzed and modified, it is time to extract results and statistics. Before that, as mentioned above, it is necessary to balance the corpus. By extracting the entire database of actors, linguists and physicists, the gender balance has not been taken into account, so that after an analysis it has been seen that the percentage of female linguists and physicists is disproportionately lower. For this reason it has been decided to balance the corpus according to the following priorities:

- Maximum possible number of linguists and female physicists possible and subsequently proportioned to the male gender.
- Maximum possible number of actresses proportioned with actors.

However, before running the software, separate files have been filtered by length according to actor, linguist or physicist as well as for male and female gender. Doing this, the number of segments and words is as balanced as possible. That is to say, it has been separated in this way so that the final corpus is balanced in number of words and segments for actors, separately for linguists, and for physicists the same. In this way, the desired balanced corpus has been obtained.

The remaining male authors who are not used for corpus balance are kept out of the balanced corpus. So additionally these remaining male authors are added to a separate unbalanced corpus so that the information can be exploited and used.

4 Results and statistics

In this chapter, the main parts of the experimental work are explained in detail, as well as the results of the final balanced corpus.

4.1 Multilingual Gender-Balanced Corpus of In-domain Wikipedia Biographies

The first step in the creation of the corpus is the parallelization of the text contents of each author in each language. As mentioned above, the information of each author extracted from Wikipedia was different according to language, so the manual work of parallelizing the same information for each language has been the most expensive and time consuming work. For this reason, 41.1% of the authors have filtered the entire database extracted with Wiki-Tailor (either due to the difficulty of parallelising the text or simply due to the textual incoherence of the content of a given author), resulting in a balanced corpus with the following characteristics:

Total number of files	1190
Number of files per language	238
Number of female files	595
Number of male files	595

Figure 4.1: General statistics.

This balanced corpus consists mainly of actors/actresses, male physicists/female physicists and male linguists/female linguists. However, in the case of linguists and physicists, the amount of female gender in the extracted database was practically null, so that in the end the proportion of the final balanced corpus was 3% linguists and 97% actors of both female and male gender.

Once all the resulting files have been obtained, the correct conversion of the files has been checked and a semi-automatic plug-in of the created software has been used to extract statistics from the final corpus, obtaining the following results. This plug-in has been created with the same programming language, in Visual Basic for Applications and what it does is map the entire folder of files indicating the number of segments that contains the total of files and the words it finds. In this last case, in the words it finds, they automatically subtract from

each file the equivalent to words that are not part of the information of each author (terms such as xml tags, segment numbers, titles, etc.).

Total number of segments	12653
Total average of segments per file	20
Total number of words	230608
Total average of words per file	387

Figure 4.2: Female specific statistics.

Total number of segments	12477
Total average of segments per file	20
Total number of words	230605
Total average of words per file	387

Figure 4.3: Male specific statistics.

Total number of segments	25130
Total average of segments per file	20
Total number of words	461213
Total average of words per file	387

Figure 4.4: Total specific statistics.

As can be seen in the results of creating the corpus by focusing on balancing the number of words according to actors and on the other hand according to linguists and physicists, it is possible to see that the balanced corpus has achieved almost one hundred percent with only a margin of error of three words. On the other hand the corpus is correctly balanced in number of segments, because is also proportional according to gender.

Obtaining these statistics it is possible to modify and improve the corpus according to the application in order to obtain more satisfactory results for the experiments carried out using this balanced corpus.

5 Conclusions and Further Research

After creating the corpus by extracting the database of actors, physicists and linguists from Wikipedia using the Wiki-Tailor tool, is it possible to obtain conclusions that would help to create a much better corpus than the one created and consequently improve the bias when applying this corpus to future applications.

In this project, a multilingual corpus of biographies extracted from Wikipedia of actors, physicists and linguists has been presented. The corpus has a defined structure where each file is identified with the original Wikipedia article ID, language, domain and genre of the author, and this is represented by xml. This structure allows the corpus to be divided into different subsets in order to evaluate the accuracy of the translation in future works.

The main characteristic of the corpus is that it is balanced both by gender and by number of words, so a balanced subset is extracted that is useful for the evaluation of MT at document level. For this reason, since we can have so many variants and experiment according to the files that make up the corpus, the results are not definitive, as they can have modifications according to the study that we want to carry out.

One aspect to keep in mind about the course of work is, as we have said before, a previous study of the statistics could have been carried out before extracting the corpus, since, as we have seen, the corpus contains a quantity of feminine genre quite inferior to the quantity of masculine genre.

On the other hand, one aspect that has hindered the optimal accomplishment of the work, has been the difference of content and information of each author depending on each language. If Wikipedia had the exact information of each author in the different languages worked, the task would have been much more effective, but having to parallelize all the contents to have the same amount of information for the five different languages, an excessive time has been dedicated. Also, in order to create a corpus with sufficient information, it has been necessary to translate sentences in different languages to have the parallel structure of the information, and there has not been enough time to carry out the post-process of checking the naturalness of the language and correcting the translation errors produced.

Choosing this way, at least, it has been possible to have the structure of a great amount of information having the parallel information in five different languages and although now it remains to carry out the process of reviewing the translation errors, the time spent is much more optimal than having to complete the corpus analyzing each author one by one in five different languages without translating each sentence in each language.

The software created has helped to be able to have the optimal characteristics to be able to

carry out the studies in a more effective way, but the manual parallelization and translation of each text file could give a margin of error to take into account that post-processing as future work will correct.

Also, it is possible to add a future development in addition to the above-mentioned post-processing correction, related to the most applicative part of this world, where we could use and improve the corpus created for different applications related to the classification of sentences by genre, the reduction of gender bias in natural language processing, etc. In addition, if this corpus was created by a each language specialist, it would really be possible to create a much better corpus, since currently a corpus has been created in five different languages where there may be a relative margin of error due to ignorance of the language and having to resort to translator.

Finally, to add that from the improvement of the created software many more corpus can be obtained to realize different applications from databases not only of Wikipedia, but of other different sources where the massive extraction of data can make difficult the manual work of all the files that form the future corpus.

6 Bibliography

[1] Kelle Webster. Marta Recasens. Vera Axelrod. Jason Baldridge. "A Balanced Corpus of Gendered Ambiguous Pronouns". Google AI Language. October 2018.

[2] Tolga Bolukbasi. Kai.Wei Chang. James Zou. Venkatesh Saligrama. Adam Kalai. "Man is to Computer Programmer as Woman is to Homemaker?Debiasing Word Embeddings. Boston University. Microsoft Research New England, 1 Memorial Drive, Cambridge, MA. May 2018.

[3] Adám Varga. Domain adaptation for multilingual neural machine translation. Master's thesis, Saarland University, August 2017.

[4] Ines Abbes. Mohamed Jemni. Towards OpenDomain CrossLanguage Question Answering. In Qatar Foundation Annual Research Conference Proceedings, March 2018.

[5] J. Holmes. M. Meyerhoff. The handbook of language and gender, volume 24. John Wiley Sons, 2008.

[6] M.O.R. Prates, P. H. C. Avelar, and L. Lamb. Assesing Gender Bias in Machine Translation – A Case Study with Google Translate. ArXiv e-prints, September 2018.

[7] Kaiji Lu. Piotr Mardziel. Fangjing Wu. Preetam Amancharla. Anupam Datta. "Gender Bias in Neural Natural Language Processing". Carneigie Mellon University. Moffiet Field, CA 94035, July 2018.

[8] J.H. Park, J. Shin, and P. Fung. Reducing Gender Bias in Abusive Language Detection. ArXiv e-prints, August 2018.

[9] Kaiji Lu. Piotr Mardziel. Fangjing Wu. Preetam Amancharla. Anupam Datta. "Gender Bias in Neural Natural Language Processing". Carneigie Mellon University. Moffiet Field, CA 94035, July 2018.

[10] Cristina España-Bonet, "WikiTailor and Multilingual Corpora from Wikipedia," Ph.D. dissertation, University of Saarland, Germany, 2018.

7 Annexes

7.1 Discarded files after balance

When parallelizing and converting more male than female text files, the corpus was not balanced, so it was decided to discard a number of male text files.

Also, there are files of female gender that have had to be discarded in order to balance correctly in terms of the proportion of professions and the number of words.

The discard has been a function of the files that have been used to create the balanced corpus, that is to say, in such a way that the same number of male and female actors, physicists and linguists result. For this reason, it has been decided to create a separate non-balanced corpus in order to be able to use this amount of data. The statistics of this corpus are as follows:

Total number of files	655
Number of files per language	131
Total number of segments	11638
Total average of segments per file	17
Total number of words	217526
Total average of words per file	332

Figure 7.1: General statistics.